

On The Inverse Geostatistical Problem of Inference on Missing Locations

Emanuele Giorgi & Peter J. Diggle
Lancaster Medical School
Lancaster University

September 12, 2014

Abstract

The standard geostatistical problem is to predict the values of a spatially continuous phenomenon, $S(x)$ say, at locations x using data $(y_i, x_i) : i = 1, \dots, n$ where y_i is the realisation at location x_i of $S(x_i)$, or of a random variable Y_i that is stochastically related to $S(x_i)$. In this paper we address the inverse problem of predicting the locations of observed measurements y . We discuss how knowledge of the sampling mechanism can and should inform a prior specification, $\pi(x)$ say, for the joint distribution of the measurement locations $X = \{x_i : i = 1, \dots, n\}$, and propose an efficient Metropolis-Hastings algorithm for drawing samples from the resulting predictive distribution of the missing elements of X . An important feature in many applied settings is that this predictive distribution is multi-modal, which severely limits the usefulness of simple summary measures such as the mean or median. We present two simulated examples to demonstrate the importance of the specification for $\pi(x)$, and analyse rainfall data from Paraná State, Brazil to show how, under additional assumptions, an empirical estimate of $\pi(x)$ can be used when no prior information on the sampling design is available.

Keywords: Geostatistics; kernel density estimation; missing locations; multi-modal distributions.

1 Introduction

Geostatistics was originally developed as a self-contained methodology for spatial prediction (e.g. Mathéron (1963)) but is now embedded as a sub-branch of spatial statistics with applications in many different disciplines. The canonical geostatistical problem is to predict the value of a spatially continuous process, $S(x)$ say, at any required location x in a region of interest $A \subset \mathbb{R}^2$, using data consisting of a set of measured values y_i at each of n locations x_i in A . A widely used geostatistical model is that the y_i are realisations of random variables $Y_i = S(x_i) + Z_i$, where Z_i are mutually independent, zero-mean Gaussian variables, and $S = \{S(x) : x \in \mathbb{R}^2\}$ is a Gaussian process (Diggle et al., 1998). Predictive inference for S is then based on the predictive distribution $[S|Y]$, where $[\cdot]$ means “the distribution of” and $Y = (Y_1, \dots, Y_n)$. Conventionally, the set of measurement locations $X = (x_1, \dots, x_n)$ is regarded either as fixed or, if stochastic, as non-informative in the sense that S and X are

stochastically independent; Diggle et al. (2010) refer to this assumption as *non-preferential* sampling.

In this paper we address the inverse problem of predictive inference for missing locations associated with a subset of the measured variables Y . Our interest in this problem was motivated by Wasser et al. (2004). They proposed a geostatistical model for data on 399 DNA samples of elephant tusks collected from 28 distinct known locations across Africa, and showed how their model could be used to infer the geographic location of a DNA sample of unknown origin, in order to identify areas of intense poaching. For their Bayesian analysis, they specified a uniform prior for sample locations over either savannah or forest regions of Africa, according to each sample's known origin and used the corresponding posterior median of each spatial coordinate as a point prediction of the unknown location. They evaluated predictive performance empirically using leave-one-out cross-validation, and found that their geostatistical approach outperformed other methods that do not use geographic information.

Our objectives in this paper are the following: to extend the approach of Wasser et al. (2004) to include multiple missing locations; to show how prior information on the sampling design can and should inform the analysis; and to highlight the unsatisfactory nature of simple statistics such as the mean or the median as summaries of the predictive distribution of missing locations.

The paper is structured as follows. In Section 2 we propose a modelling framework for predictive inference on multiple missing locations, incorporating prior knowledge about the sampling design. In Section 3 we show how numerical quadrature can be used for predictive inference on a single unknown location, and propose an MCMC algorithm for sampling from the joint predictive distributions of multiple unknown locations. In both cases we assume that parameter values are known, which in practice corresponds to ignoring the uncertainty parameter estimates. In Section 4, we describe two simulated examples to illustrate the limitations of simple summary statistics of the predictive distribution for the missing locations. We also report an analysis of rainfall data from Paraña State, Brazil, where we use a non-parametric density estimate as an empirical prior for X . Section 5 is a concluding discussion.

2 Model formulation

2.1 Measurement data

We adopt a standard geostatistical model,

$$Y_i = \mu_i + S(x_i) + Z_i : i = 1, \dots, n, \quad (1)$$

where $S = \{S(x) : x \in \mathbb{R}^2\}$ is a zero-mean stationary Gaussian process with variance σ^2 and correlation function $\rho(x, x'; \phi)$ indexed by the parameter ϕ , and the Z_i are mutually independent $N(0, \tau^2)$ variates. Equivalently, the Y_i are conditionally independent given $\{S(x) : x \in \mathbb{R}^2\}$, with $[Y_i | S(x_i)] = N(\mu_i + S(x_i), \tau^2)$. Write X^* for a set of unknown locations at which measurements Y^* have been made, $X = (\tilde{X}, X^*)$ and $Y = (\tilde{Y}, Y^*)$; the observed quantities are \tilde{X} and Y . For any set \mathcal{X} of points $x \in \mathbb{R}^2$ write $S(\mathcal{X}) = \{S(x) : x \in \mathcal{X}\}$; hence, $S(X) = (S(\tilde{X}), S(X^*))$. Assume that X and S are stochastically independent. The joint distribution of X , Y and S is then

$$\begin{aligned} [X, Y, S] &= [S] [X] [Y | S, X] \\ &= [S] [\tilde{X}] [X^* | \tilde{X}] [\tilde{Y} | S(\tilde{X})] [Y^* | S(X^*)], \end{aligned} \quad (2)$$

Our assumption that the sampling design is non-preferential allows a straightforward marginalisation of (2) to give

$$\begin{aligned} [X, Y] &= [\tilde{X}] [X^*|\tilde{X}] [\tilde{Y}|X] [Y^*|\tilde{Y}, X] \\ &= [\tilde{X}] [X^*|\tilde{X}] [\tilde{Y}|\tilde{X}] [Y^*|\tilde{Y}, X^*] \\ &= [\tilde{X}] [X^*|\tilde{X}] [Y|X] \end{aligned} \quad (3)$$

where $[Y|X]$ is a multivariate Gaussian distribution with mean vector $\mu = (\mu_1, \dots, \mu_n)$ and covariance matrix Σ with diagonal elements $\sigma^2 + \tau^2$ and off-diagonal elements $\sigma^2 \rho(x_i, x_j; \phi)$.

2.2 Sampling design

Depending on the problem under investigation, the set of sampling locations might be the result of a natural process, for example the locations of nests in a colony of birds. Alternatively, they might be obtained by using a random or regular lattice designs, as it is often the case for household surveys or agricultural field trials, respectively. Knowledge of the underlying process generating the sampling locations should then be incorporated into the specification of the distribution $[\tilde{X}]$. We now briefly outline some approaches to this specification, and propose a non-parametric approach that can be used when information on the underlying sampling process is limited.

One approach would be to model $[\tilde{X}]$ as an inhomogeneous Poisson process over the region of interest $A \subseteq \mathbb{R}^2$, with intensity

$$\lambda(x) = d(x)^\top \beta, \quad (4)$$

where $d(x)$ is a p -dimensional vector of spatial covariates, such as population density in the case of a randomised household survey, and β is the associated vector of regression coefficients.

When no information on the sampling design is available a non-informative uniform distribution could be used, hence $\lambda(x) = \lambda$ for all $x \in A$. An alternative approach is to estimate the intensity $\lambda(x)$ from the data, using a kernel method. Let x_1 and x_2 denote the coordinates of the horizontal and vertical axes for a given point $x \in \mathbb{R}^2$. Then, the kernel density estimate of $\pi(x)$, i.e. the marginal density function of any component of X , based on the observed locations X^* is given by

$$\hat{\pi}(x) = \frac{1}{n} \sum_{i=1}^n K_H(x_1 - x_{1i}; x_2 - x_{2i}), \quad (5)$$

where $K_H(\cdot; \cdot)$ is a bivariate kernel with symmetric and positive definite 2 by 2 smoothing matrix H . If we choose a Gaussian kernel, then H is the variance matrix of a bivariate Gaussian density and

$$K_H(x_1 - x_{1i}; x_2 - x_{2i}) = \frac{1}{\sqrt{2\pi}|H|^{1/2}} \exp \left\{ -\frac{1}{2}(x - x_i)^\top H^{-1}(x - x_i) \right\}.$$

The elements of H can be estimated by optimising an estimate of the mean-square-error (Breman & Diggle, 1989). Alternatively, if we assume that X is an independent random sample from a bivariate Gaussian distribution, the optimal H in the sense of minimising the integrated mean-square-error is $H = n^{-1/6}V$ (Lucy et al., 2002), where V is the sample covariance matrix.

Finally, a common practice in geostatistical investigations is to choose locations that are more regularly distributed over A than would be a realisation of a homogeneous Poisson process. In this case, a more appropriate prior for $[X]$ would be an inhibitory point process (Diggle, 2013, pages 110-111); we give an example in Section 4.2.

3 Computational details

3.1 One unknown location

If X^* consists of a single unknown location, numerical quadrature can be used for efficient computation of the predictive distribution $[X^*|\tilde{X}, \tilde{Y}, Y^*]$. Let $W = (w_1, \dots, w_N)$ be a grid of spatial points in the region of interest A and let $\pi(z)$ denote the density function of Z . It follows from (3) that

$$\pi(x^*|\tilde{x}, \tilde{y}, y^*) \propto \pi(x^*|\tilde{x})\pi(y^*|\tilde{y}, x^*). \quad (6)$$

By treating the Gaussian process S as constant within each grid cell, we approximate the above density function by

$$\pi(x^*|\tilde{x}, \tilde{y}, y^*) \approx h(w^*) = \frac{\pi(w^*|\tilde{w}, \tilde{y}, y^*)}{\sum_{i=1}^N \pi(w_i|\tilde{w}, \tilde{y}, y^*)}, \quad (7)$$

where w^* and the elements of \tilde{w} are the grid points closest to x^* and to the corresponding elements of \tilde{x} , respectively. Summaries of the predictive distribution, such as mean, mode and component-wise median, can then be approximately computed through $h(\cdot)$. Additionally, high density regions of coverage α can also be obtained as

$$\left\{ w_i \in W : l \in \mathbb{R}, \sum_{i:h(w_i) > l} h(w_i) = 1 - \alpha \right\}.$$

3.2 Multiple unknown locations

When there is more than one unknown location, the numerical solution is no longer feasible. Instead, we have developed an MCMC algorithm that takes account of the presence of the multiple modes that typically characterize the density function of $[X^*|\tilde{X}, \tilde{Y}, Y^*]$, each mode corresponding to a location where the absolute difference between a value of Y^* associated with an unknown location and an observed value of \tilde{Y} is small. At each iteration of the MCMC, and with a pre-specified probability, the algorithm proposes a draw from a mixture distribution with a mode centred on each observed location.

Let n^* and \tilde{n} denote the number of unknown and known locations, respectively; we propose the following Metropolis-Hastings algorithm to simulate from the target density given by (6).

1. Initialize $x_{\text{curr}}^* = x_0^*$.
2. Propose a new value x_{prop}^* as follows. For $i = 1, \dots, n^*$:
 - with probability p perform a random walk by proposing a value from a bivariate Gaussian distribution with mean the i -th element of x_{curr}^* and covariance matrix $h_1^2 I$, where $h_1 > 0$ and I is a 2 by 2 identity matrix;

- with probability $1 - p$ sample a data point x_j uniformly from the set of observed locations and propose a value from a bivariate Gaussian distribution with mean x_j and covariance matrix $h_2^2 I$, where $h_2 > 0$.

3. Accept $x_{\text{prop.}}^*$ with probability

$$\min \left\{ 1, \frac{\pi(x_{\text{prop.}}^* | \tilde{x}, \tilde{y}, y^*) q(x_{\text{curr.}}^* | x_{\text{prop.}}^*)}{\pi(x_{\text{curr.}}^* | \tilde{x}, \tilde{y}, y^*) q(x_{\text{prop.}}^* | x_{\text{curr.}}^*)} \right\},$$

where

$$q((w_1, \dots, w_{n^*}) | (z_1, \dots, z_{n^*})) = \prod_{i=1}^{n^*} \left[\frac{p}{h_1} f(\|w_i - z_i\|/h_1) + \frac{1-p}{\tilde{n}h_2} \sum_{j=1}^{\tilde{n}} f(\|w_i - \tilde{x}_j\|/h_2) \right]. \quad (8)$$

In 8, $f(\cdot)$ is the density function of a standard Gaussian distribution and $\|\cdot\|$ is the Euclidean distance.

4. Repeat 2 and 3 for the desired number of iterations.

In this algorithm, the standard deviations h_1 , h_2 and p should be tuned manually via pilot runs.

4 Examples

4.1 Simulated data from a homogeneous Poisson process

Our first example highlights the difficulty of summarizing the multi-modal distribution $[X^* | \tilde{X}, \tilde{Y}, Y^*]$ by a single measure of location. We simulated two data sets of size 201, using an isotropic exponential correlation function for the Gaussian process $S(x)$. The parameter values were set as $\mu = 0$, $\sigma^2 = 1$, $\tau^2 = 0.1$, and $\phi = 0.1$ or $\phi = 0.5$. From each of the two simulated data-sets we treated one of the 201 locations, chosen at random, as unknown. Locations were generated uniformly in the unit square, hence $\pi(x^* | \tilde{x}) = 1$ for x^* lies in the unit square and 0 otherwise.

Figures 1(a)-(b) and 1(c)-(d) show the results for $\phi = 0.1$ and $\phi = 0.5$, respectively. The multi-modality of the distribution is indicated by numerous black patches of high density, that become more spread when $\phi = 0.5$. In Figure 1(d), the mean lies outside the 95% highest density region, whilst the mode, although not having that unpleasant feature, has only slightly larger predictive density than other local modes.

4.2 Simulated data from an inhibitory point process

We now consider the case when $[X]$ is a simple sequential inhibition process on the unit square. Denote by $\delta > 0$ the minimum permissible distance between any two locations. A sample from X is obtained by a sequential sampling of points from a 100 by 100 regular lattice, where each new location X_{i+1} given $\{X_j = x_j, j = 1, \dots, i\}$ is generated uniformly on the intersection of the unit square with $\{x \in [0, 1]^2 : \|x - x_j\| \geq \delta, j = 1, \dots, i\}$.

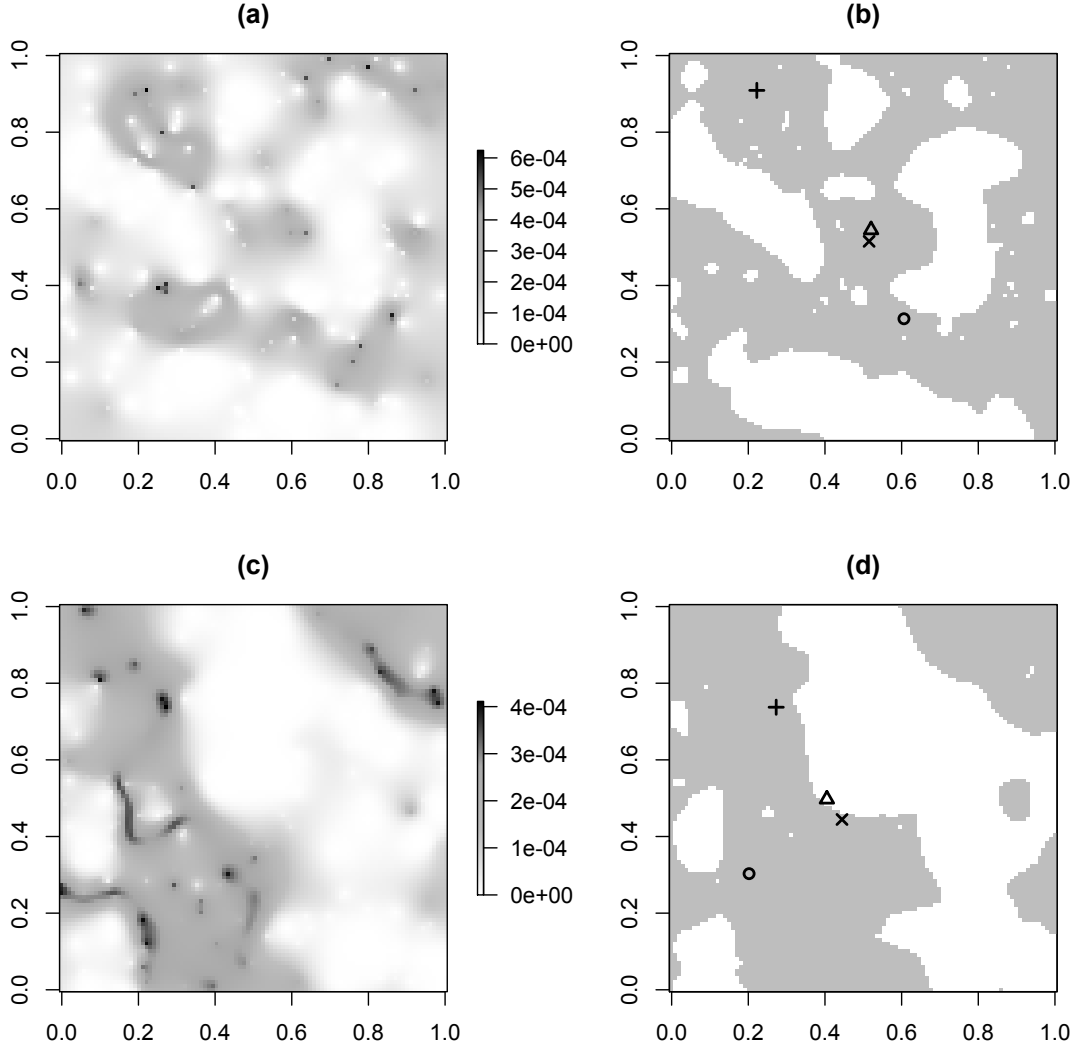


Figure 1: (a)-(c) images of the predictive distribution of one missing locations, obtained from 200 hundred locations generated uniformly over the unit square and simulated data from a Gaussian process with $\phi = 0.1$ (a) and $\phi = 0.5$ (c); (b)-(d) images of 95% high density regions for the predictive distributions in (a) and (c), respectively, showing the true location (o), the mean (\triangle), the mode (+) and the componentwise median (\times).

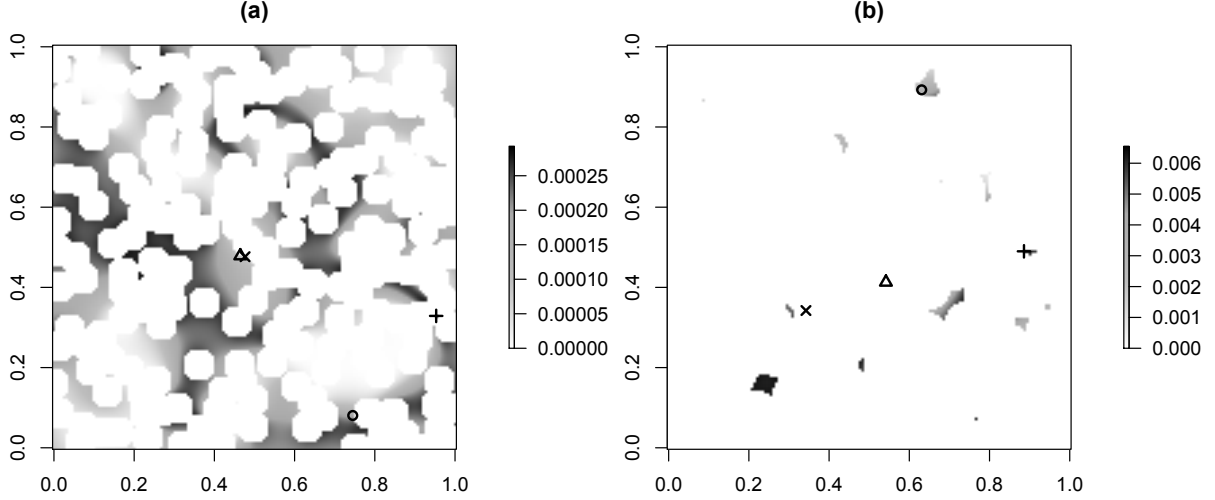


Figure 2: (a)-(b) images of the predictive distribution of one missing locations, obtained from 200 hundred locations generated from an inhibitory point process with minimum distance between locations $\delta = 0.04$ (a) and $\delta = 0.06$ (b); the true location (o), the mean (Δ), the mode (+) and the component-wise median (\times) are also shown.

We simulated two datasets of 201 observations with model parameter values set as in the previous example but fixing $\phi = 0.1$ and letting δ take values 0.04 and 0.06; again, only one location is treated as unknown. The resulting density of the predictive distribution is shown in Figure 2. As expected, the extent of the predictive distribution is reduced by imposition of the minimum permissible distance δ . In particular, for $\delta = 0.06$ in Figure 2(b), only a few small, disjoint regions remain as admissible for the missing location; note also that the mean and the median, in this specific example, lie outside the support of the predictive distribution.

4.3 Rainfall data from Paraná State in Brazil

We now consider a data set, previously analysed by Diggle & Ribeiro (2002), on average rainfall over different years for the period May-June (dry-season) recorded in 143 recording stations throughout Paraná State, Brazil. Data locations are reported in Figure 3, three of which, denoted by triangles, were randomly selected and treated as unknown.

Using the 140 observations with known locations, we first fitted the model $Y_i = \mu + S(x_i) + Z_i$, where $S(x_i)$ is an isotropic Gaussian process with variance σ^2 and Matérn (1986) correlation function

$$\rho(u; \phi, \kappa) = \{2^{k-1} \Gamma(\kappa)\}^{-1} (u/\phi)^\kappa K_\kappa(u/\phi), u > 0,$$

where u is distance between two locations, $K_\kappa(\cdot)$ denotes the modified Bessel function of the second kind of order $\kappa > 0$ and $\phi > 0$ is a scale parameter. The maximum likelihood estimates are: $\hat{\mu} = 251.539$, $\hat{\sigma}^2 = 9422.807$, $\hat{\tau}^2 = 479.074$, $\hat{\phi} = 200.004$ and $\hat{\kappa} = 1.913$. As also indicated by the resulting theoretical semi-variogram in Figure 3(b), the data show evidence of long-range spatial correlation: the most distant pair of observations at 619.492 km from each other have an estimated correlation of about 0.25.

In order to make inference on the three unknown locations, we model $[X^*]$ in two different ways: 1) $\pi(x^*|\tilde{x})$ is uniform over the square $[71.978, 846.652] \times [13.887, 518.441]$; 2) $\pi(x^*|\tilde{x})$ is estimated from the data using kernel density estimation as described in Section 2. The resulting non-parametric estimate of $\pi(x^*|\tilde{x})$ is shown in Figure 3(a); note that the boundaries of Paraná State play no part in the analysis but are displayed only to add context.

In both scenarios, using the MCMC algorithm described in Section 3, we obtained 10000 samples from the predictive distribution $[X^*|X, Y, Y^*]$, iterating 110000 times and retaining every 10th sample after a burn-in of 10000 samples; we set $h_1 = 70$, $h_2 = 31$ and $p = 0.5$. The autocorrelogram plots for the vertical and horizontal coordinates in Figure 4 suggest rapid convergence of the MCMC algorithm. The resulting predictive distributions for one of the three missing location, specifically $x^* = (692.545, 170.875)$, are shown in Figure 5 with corresponding high density regions. Panels (b) and (d) show that the high density regions obtained using a uniform distribution have a much wider extent than in the second case and are little informative on the possible positioning of x^* . However, when using the kernel density estimate $\hat{\pi}(x)$, the true location is only contained in the 95% high density region; indeed, as indicated in Figure 3(a), this point lies in a region where $\hat{\pi}(x)$ is relatively small.

5 Discussion

In this paper we have proposed a methodology that allows to make inference on unknown locations of geostatistical data. We have developed efficient quadrature and MCMC algorithms for sampling from the predictive distribution $[X^*|\tilde{X}, \tilde{Y}, Y^*]$ for single and multiple unknown locations X^* , respectively, and applied these to two simulated data-sets and to rainfall data from Paraná state, Brazil, with three unknown locations. In other examples not shown, we found that the MCMC algorithm continued to mix well when there are more unknown locations.

The conjunction of a sufficiently large data-set and many unknown locations would increase the computational burden of the MCMC algorithm. A computationally more efficient, but approximate procedure would be to use numerical quadrature in turn for each of the missing locations.

As shown in Section 4.1, the distribution $[X^*|\tilde{X}, \tilde{Y}, Y^*]$ is often characterized by multiple modes and disjoint high density regions, in which case commonly used indices, such as the mean and the median, are misleading summaries.

We have shown that the use of a homogeneous Poisson process prior $[X^*]$ may result in a very diffuse predictive distribution with widespread regions of high density. The use of a kernel density estimate for $\pi(x^*|\tilde{x})$, based on the set of observed locations is useful when the empirical distribution of \tilde{x} is spatially heterogeneous and the conditional distribution of the unknown locations X^* is expected to follow the same pattern. Conversely, an inhibitory process for $[X^*|\tilde{X}]$ is more appropriate when the context suggests that the complete set of locations $X = (\tilde{X}, X^*)$ is likely to show some degree of spatial regularity.

We have not considered incorporation of observations with unknown locations into a likelihood for parameter estimation. However, this can be pursued by noticing that, starting from (3), the likelihood is given by

$$\pi(\tilde{y}, y^*) \propto \pi(\tilde{y}|\tilde{x}) \times \int_{\mathbb{R}^{2n^*}} \pi(x^*|\tilde{x}) \pi(y^*|\tilde{y}, x^*) dx^*,$$

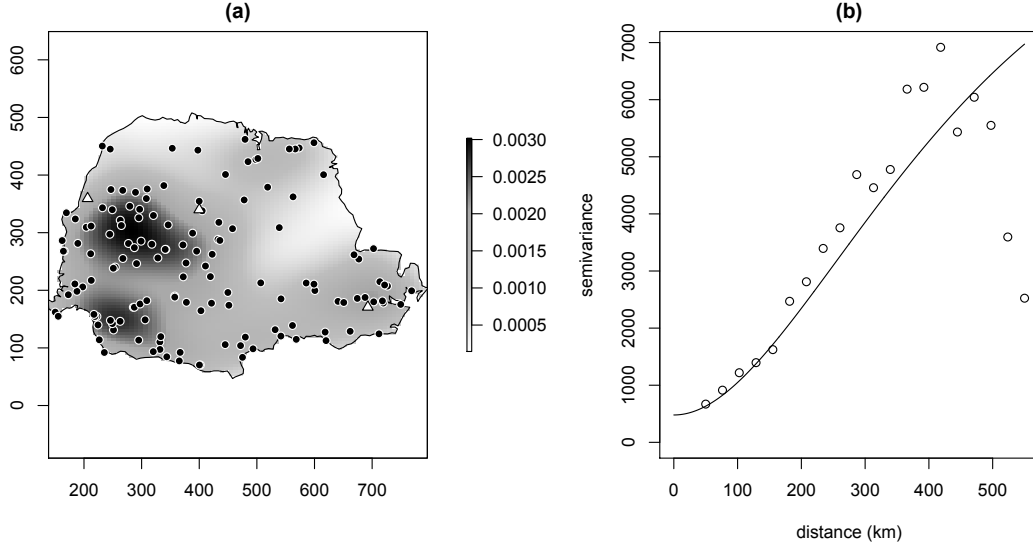


Figure 3: (a) observed (solid circles) and missing (triangles) geographical coordinates of the recording station, with a kernel density estimate of $\pi(x^*)$ on the background; (b) empirical (points) and theoretical (solid line) semi-variogram.

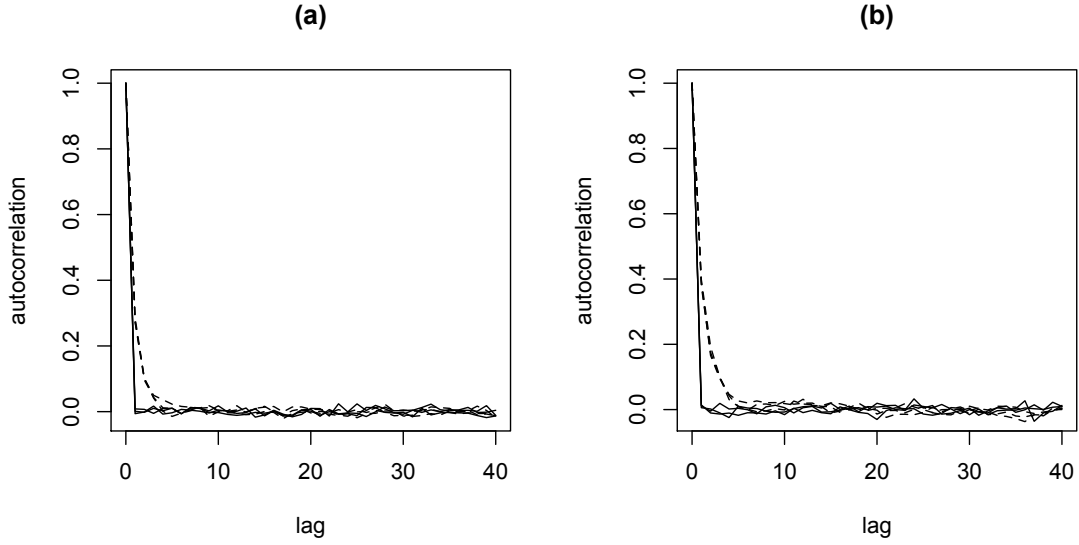


Figure 4: autocorrelogram of the 10000 samples from $[X^*|X, Y, Y^*]$ for the horizontal (a) and (b) vertical coordinates, using a uniform distribution (solid lines) and a kernel density estimate (dashed lines) for $\pi(x^*)$.

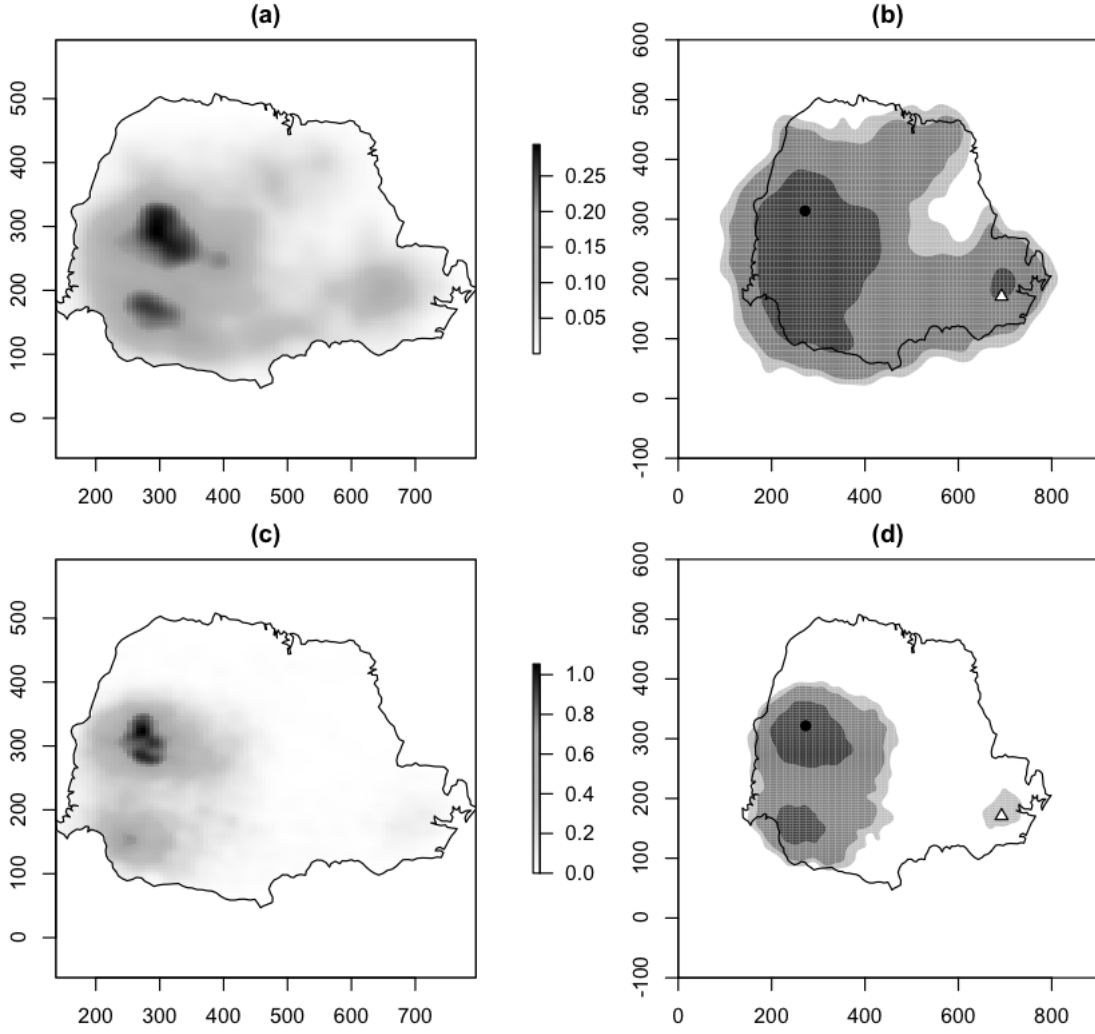


Figure 5: (a),(c) images of the predictive distribution $\pi(x^*|\tilde{x}, \tilde{y}, y^*)$ for the missing location $x^* = (692.545, 170.875)$. (a) $\pi(x^*)$ uniform over the square $[71.978, 846.652] \times [13.887, 518.441]$. (c) kernel density estimate for $\pi(x^*)$; (b),(d) 95%, 90% and 50% highest density regions for the predictive distributions in (a) and (c), respectively, indicating the mode (solid circle) and the true location x^* (triangle).

where the first factor is the standard likelihood function obtained from data at known locations. The second factor is an intractable integral of dimension twice the number of unknown locations, n^* say. As discussed in Section 3, either numerical quadrature or MCMC can be used to evaluate the integral according to the value of n^* . More efficient Monte Carlo techniques based on importance sampling or the EM algorithm could also be considered. For Bayesian inference, we would need additionally to specify a prior for the model parameters. It would then be interesting to determine under what circumstances the incorporation of observations with missing locations into the model-fitting process leads to materially better parameter estimates. We conjecture that the additional information will be small unless the prior for $[X^*|\tilde{X}]$ is highly informative.

Finally, we have assumed throughout that $[X]$, and hence $[X^*|\tilde{X}]$, is stochastically independent of $[S]$. Extensions of the modelling framework that allow for stochastic dependence between X and the latent process S would add to the predictability of unknown locations X^* but would also complicate the fitting of the model. One of a number of possibilities would be to model X as a Cox Process with intensity stochastically dependent on S , as in Diggle, Menezes & Su (2010).

References

- BREMAN, M. & DIGGLE, P. J. (1989). Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society, Series B* **51**, 81–92.
- DIGGLE, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Chapman and Hall/CRC, 3rd ed.
- DIGGLE, P. J., MENEZES, R. & SU, T. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society, Series C* **59**, 191–232.
- DIGGLE, P. J. & RIBEIRO, P. J. (2002). Bayesian inference in Gaussian model-based geostatistics. *Geographical and Environmental Modelling* **6**, 129–146.
- DIGGLE, P. J., TAWN, J. A. & MOYEED, R. A. (1998). Model-based geostatistics (with discussion). *Applied Statistics* **47**, 299–350.
- LUCY, D., AYKROYD, R. & POLLARD, A. (2002). Non-parametric calibration for age estimation. *Applied Statistics* **51**, 183–196.
- MATÉRN, B. (1986). *Spatial Variation*. Springer, Berlin, 2nd ed.
- MATHÉRON, G. (1963). Principles of geostatistics. *Economic Geology* **58**, 1246–1266.
- WASSER, S. K., SHEDLOCK, A. M., COMSTOCK, K. E., OSTRANDER, E. A. & MUTAYOBA, B. (2004). Assigning African elephant DNA to geographic region of origin: applications to the ivory trade. *Proceedings of the National Academy of Science (USA)* **101**, 14847–14852.